



UNIVERSIDAD DE MONTERREY

División de Ingeniería y Tecnología

Inteligencia Artificial

K-Nearest-Neighbor

Gabriel Aldahir López Soto **#552543**

Dr. Andrés Hernández Gutiérrez

San Pedro Garza García, N.L. 13 de mayo, 2020

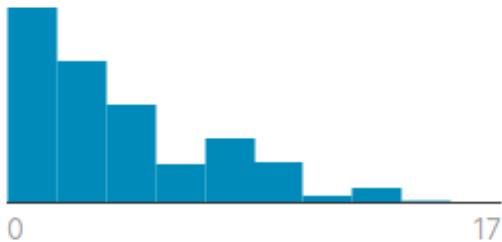
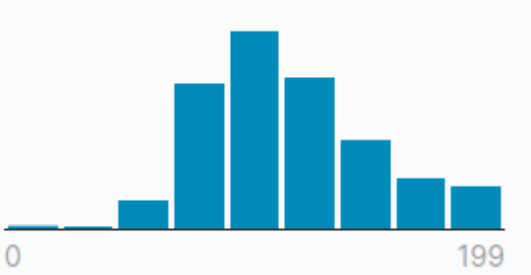
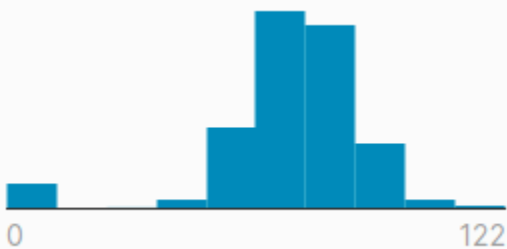
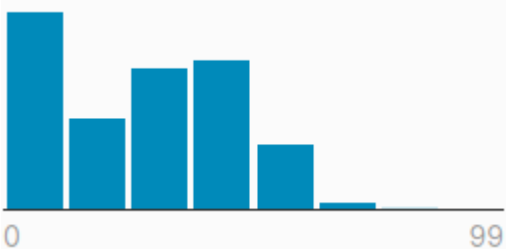
K-Nearest-Neighbor

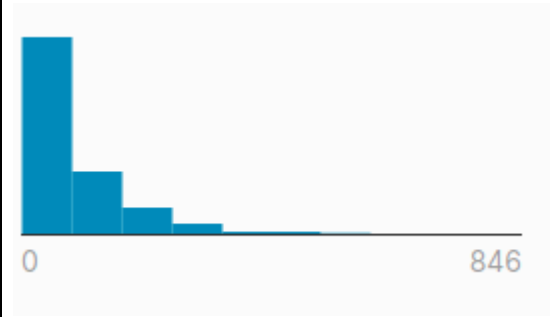
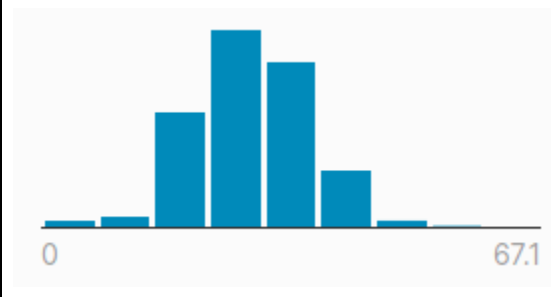
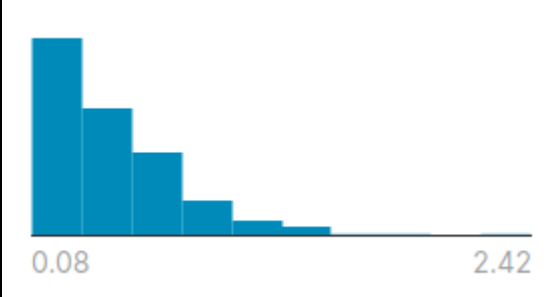
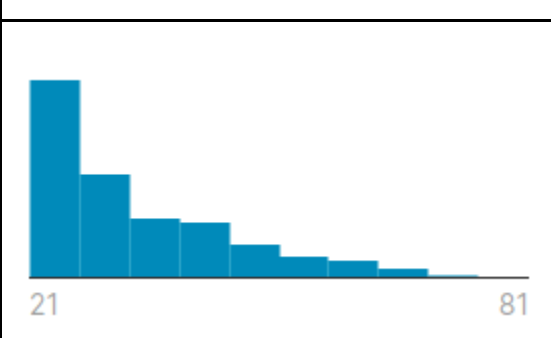
Este algoritmo o modelo de clasificación está basado en posiciones, es decir clasifica de acuerdo con cercanía, es decir el algoritmo clasifica los datos en el grupo que tenga los k vecinos más cerca. Al igual que el trabajo pasado este modelo es de clasificación binaria, 0 o 1 el cual 0 significa que no pertenece a la clase y 1 que si pertenece a la clase. Este modelo calcula el dato (x_0) con respecto al conjunto de datos y saca su distancia Euclidiana, se ordenan las distancias y se calcula el número de frecuencia de los resultados para determinar si pertenece o no al grupo.

Predicción de Diabetes

La diabetes es un tema muy importante porque tengo familiares y amigos que padecen esta enfermedad que ataca sin piedad a las personas, si bien la diabetes y cuando tu cuerpo no produce ni consume la hormona insulina lo que provoca un exceso de azúcar en tu sangre. En este trabajo los datos fueron obtenidos del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. Estos datos tienen como objetivo predecir de forma diagnóstica si un paciente tiene diabetes o no, basándose en ciertas mediciones incluidas en el conjunto de datos. Las variables por utilizar son el número de embarazos que ha tenido la paciente, su IMC, nivel de insulina, edad, etc.

Características:

Características	Histograma
Embarazos (número de veces de embarazo)	 <p>0 17</p>
Glucosa (concentración de glucosa en plasma 2 horas en una prueba de tolerancia oral a la glucosa)	 <p>0 199</p>
Presión arterial (presión arterial diastólica (mmHg))	 <p>0 122</p>
Espesor de la piel (espesor del pliegue de la piel del tríceps (mm))	 <p>0 99</p>

Insulina (insulina sérica de 2 horas (mu U / ml))	 <p>0 846</p>
IMC (índice de masa corporal (peso en kg / (altura en m) ^ 2))	 <p>0 67.1</p>
DiabetesPedigreeFunction (función de pedigrí de la diabetes)	 <p>0.08 2.42</p>
Edad en años	 <p>21 81</p>

Discusión

Para el algoritmo se necesitamos leer el data set que contiene 8 características antes mencionadas y una columna que indica el resultado, es decir, si es o no diabético, si bien dentro del data set los datos son muy volátiles en rangos y son características que no tienen tanta relación una de otra. Y un punto que vi es que el tamaño de números de embarazos es muy grande en la india.

Para la construcción se ocupó leer la información y separar las características en una variable aparte y sobre eso aplicar un entrenamiento de 95% y prueba de 5% todos a su vez son aleatorios.

Después recorrimos todos los datos de pruebas y en cada conjunto compuesto de 8 características calculábamos si era o no diabético, para eso calculábamos la distancia euclidiana, después ordenábamos los datos y sacábamos los k más cercanos a 0, con base a eso mediamos la frecuencia de los resultados de los k primeros datos y obteníamos si era o no diabético. Con esa predicción construimos una matriz de confusión en la cual obtendremos los falsos negativos (FN), falsos positivos (FP), verdaderos positivos (TP), verdaderos negativos (TN). Y al final con base a eso podremos obtener las métricas de exactitud, precisión, sensibilidad, especifico y puntuación F1.

Pregnancies	Glucose	BloodPressure	Skin Thickness	Insulin	BMI	Diabetes.Ped.Fun.	Age	Pb. Diabetes	Pb. NO Diabetes
4.0	116.0	72.0	12.0	87.0	22.1	0.463	37.0	0.9	0.1
9.0	106.0	52.0	0.0	0.0	31.2	0.38	42.0	0.3	0.7
10.0	122.0	68.0	0.0	0.0	31.2	0.258	41.0	0.3	0.7
4.0	148.0	60.0	27.0	318.0	30.9	0.15	29.0	0.8	0.2
0.0	128.0	68.0	19.0	180.0	30.5	1.391	25.0	0.8	0.2
1.0	100.0	72.0	12.0	70.0	25.3	0.658	28.0	0.9	0.1
3.0	148.0	66.0	25.0	0.0	32.5	0.256	22.0	0.8	0.2
5.0	130.0	82.0	0.0	0.0	39.1	0.956	37.0	0.4	0.6
0.0	93.0	60.0	0.0	0.0	35.3	0.263	25.0	0.7	0.3
1.0	112.0	80.0	45.0	132.0	34.8	0.217	24.0	1.0	0.0
2.0	95.0	54.0	14.0	88.0	26.1	0.748	22.0	0.9	0.1
7.0	114.0	64.0	0.0	0.0	27.4	0.732	34.0	0.5	0.5
2.0	129.0	0.0	0.0	0.0	38.5	0.304	41.0	0.1	0.9
2.0	146.0	76.0	35.0	194.0	38.2	0.329	29.0	0.8	0.2
0.0	105.0	90.0	0.0	0.0	29.6	0.197	46.0	0.4	0.6
5.0	162.0	104.0	0.0	0.0	37.7	0.151	52.0	0.4	0.6
8.0	188.0	78.0	0.0	0.0	47.9	0.137	43.0	0.2	0.8
8.0	112.0	72.0	0.0	0.0	23.6	0.84	58.0	0.9	0.1
2.0	105.0	75.0	0.0	0.0	23.3	0.56	53.0	0.4	0.6
5.0	0.0	80.0	32.0	0.0	41.0	0.346	37.0	0.9	0.1
2.0	112.0	75.0	32.0	0.0	35.7	0.148	21.0	0.8	0.2
3.0	129.0	92.0	49.0	155.0	36.4	0.968	32.0	0.4	0.6
13.0	152.0	90.0	33.0	29.0	26.8	0.731	43.0	0.4	0.6
4.0	129.0	60.0	12.0	231.0	27.5	0.527	31.0	0.9	0.1
3.0	170.0	64.0	37.0	225.0	34.5	0.356	30.0	0.3	0.7
5.0	96.0	74.0	18.0	67.0	33.6	0.997	43.0	0.4	0.6
0.0	146.0	70.0	0.0	0.0	37.9	0.334	28.0	0.7	0.3
0.0	129.0	80.0	0.0	0.0	31.2	0.703	29.0	0.7	0.3
6.0	154.0	78.0	41.0	140.0	46.1	0.571	27.0	0.4	0.6
2.0	110.0	74.0	29.0	125.0	32.4	0.698	27.0	0.6	0.4
2.0	56.0	56.0	28.0	45.0	24.2	0.332	22.0	0.9	0.1
2.0	158.0	90.0	0.0	0.0	31.6	0.805	66.0	0.5	0.5
4.0	137.0	84.0	0.0	0.0	31.2	0.252	30.0	0.7	0.3
4.0	132.0	0.0	0.0	0.0	32.9	0.302	23.0	0.4	0.6
4.0	90.0	0.0	0.0	0.0	28.0	0.61	31.0	0.8	0.2
5.0	108.0	72.0	43.0	75.0	36.1	0.263	33.0	0.9	0.1
8.0	126.0	74.0	38.0	75.0	25.9	0.162	39.0	0.6	0.4
4.0	127.0	88.0	11.0	155.0	34.5	0.598	28.0	0.7	0.3

```

Confusion Matrix
-----
TP  9  |  FP  7
-----
FN  4  |  TN  18
-----
Performance Metrics
-----
Accuracy:      0.7105263157894737
Precision:     0.5625
Recall:        0.6923076923076923
Specificity:   0.72
F1:            0.6206896551724138

```

Resultados

Dentro de los resultados vemos primeramente que si aplicamos un escalado a las características podremos ocupar menos espacio de procesamiento, cada vez que se corre el programa obtenemos resultados diferentes debido a que se toman diferentes datos de entrenamiento y prueba, además que es importante determinar el número de k vecinos que se quiere obtener, si bien entre más k vecinos llega a existir un ligero cambio en la precisión.

k	Accuracy	Precision	Recall	Specificity	F-1 Score
5	0.684210526	0.555555556	0.3846154	0.84	0.454545455
10	0.710526316	0.714285714	0.3571429	0.916666667	0.476190476
20	0.868421053	0.818181818	0.75	0.923076923	0.782608696

Conclusiones

Al tener muchas características que no sean importantes puede provocar que el algoritmo no sea preciso y aparte los datos tienen rangos muy distintos por lo que fue opcional aplicar el escalado de características para mejorar notablemente el tiempo de procesamiento.

Si bien los datos obtenidos fueron los TN, TP, pero por cada corrida variaba mucho los FP y FN, lo que demuestra que no tiene una consistente precisión por lo que se concluye que existe muchas características que hacen ruido en los resultados afectando el modelo.

“Doy mi palabra que he realizado esta tarea con Integridad Académica.”